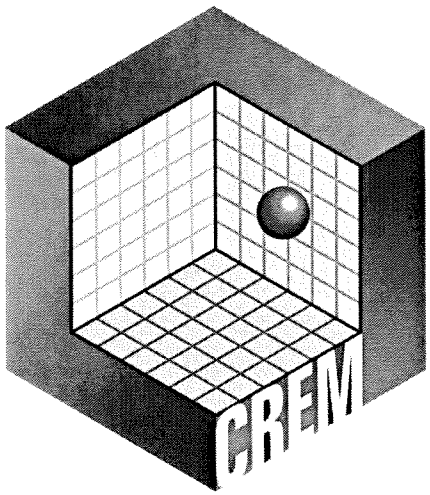
EPA/100/K-09/003 | March 2009
www.epa.gov/crem

United States
Environmental Protection
Agency

Guidance on the Development, Evaluation, and Application of Environmental Models



Office of the Science Advisor
Council for Regulatory Environmental Modeling

4. Model Evaluation

Summary of Recommendations for Model Evaluation

appropriately used to inform a decision.

- Model evaluation addresses the soundness of the science underlying a model, the quality and quantity of available data, the degree of correspondence with observed conditions, and the appropriateness of a model for a given application.
- Recommended components of the evaluation process include: (a) credible, objective peer review; (b) QA project planning and data quality assessment; (c) qualitative and/or quantitative model corroboration; and (d) sensitivity and uncertainty analyses.
- Quality is an attribute of models that is meaningful only within the context of a specific model application. Determining whether a model serves its intended purpose involves in-depth discussions between model developers and the users responsible for applying for the model to a particular problem.
- Information gathered during model evaluation allows the decision maker to be better positioned to formulate decisions and policies that take into account all relevant issues and concerns.

4.1 Introduction

> *Models will always be constrained by computational limitations, assumptions and knowledge gaps. They can best be viewed as tools to help inform decisions rather than as machines to generate truth or make decisions. Scientific advances will never make it possible to build a perfect model that accounts for every aspect of reality or to prove that a given model is correct in all aspects for a particular regulatory application. These characteristics...suggest that model evaluation be viewed as an integral and ongoing part of the life cycle of a model, from problem formulation and model conceptualization to the development and application of a computational tool.*
>
> — *NRC Committee on Models in the Regulatory Decision Process (NRC 2007)*

The natural complexity of environmental systems makes it difficult to mathematically describe all relevant processes, including all the intrinsic mechanisms that govern their behavior. Thus, policy makers often rely on models as tools to approximate reality when making decisions that affect environmental systems. The challenge facing model developers and users is determining when a model, despite its uncertainties, can be appropriately used to inform a decision. Model evaluation is the process used to make this determination. In this guidance, model evaluation is defined as *the process used to generate information to determine whether a model and its analytical results are of a quality sufficient to serve as the basis for a decision*. Model evaluation is conducted over the life cycle of the project, from development through application.

Box 5: Model Evaluation Versus Validation Versus Verification

Model evaluation should not be confused with model validation. Different disciplines assign different meanings to these terms and they are often confused. For example, Suter (1993) found that among models used for risk assessments, misconception often arises in the form of the question "Is the model valid?" and statements such as "No model should be used unless it has been validated." Suter further points out that "validated" in this context means (a) proven to correspond exactly to reality or (b) demonstrated through experimental tests to make consistently accurate predictions.

Because every model contains simplifications, predictions derived from a model can never be completely accurate and a model can never correspond exactly to reality. In addition, "validated models" (e.g., those that have been shown to correspond to field data) do not necessarily generate accurate predictions of reality for multiple applications (Beck 2002a). Thus, some researchers assert that no model is ever truly "validated"; models can only be invalidated for a specific application (Oreskes et al. 1994). Accordingly, this guidance focuses on process and techniques for *model evaluation* rather than model validation or invalidation.

"Verification" is another term commonly applied to the evaluation process. However, in this guidance and elsewhere, model verification typically refers to model code verification as defined in the model development section. For example, the NRC Committee on Models in the Regulatory Decision Process (NRC 2007) provides the following definition:

> *Verification* refers to activities that are designed to confirm that the mathematical framework embodied in the module is correct and that the computer code for a module is operating according to its intended design so that the results obtained compare favorably with those obtained using known analytical solutions or numerical solutions from simulators based on similar or identical mathematical frameworks.

In simple terms, model evaluation provides information to help answer four main questions (Beck 2002b):

1. How have the principles of sound science been addressed during model development?
2. How is the choice of model supported by the quantity and quality of available data?
3. How closely does the model approximate the real system of interest?
4. How does the model perform the specified task while meeting the objectives set by QA project planning?

These four factors address two aspects of model quality. The first factor focuses on the intrinsic mechanisms and generic properties of a model, *regardless of the particular task to which it is applied*. In contrast, the latter three factors are evaluated in the context of the use of a model *within a specific set of conditions*. Hence, it follows that model quality is an attribute that is meaningful only within the context of a *specific model application*. A model's quality to support a decision becomes known when information is available to assess these factors.

The NRC committee recommends that evaluation of a regulatory model continue throughout the life of a model and that an evaluation plan could:

- Describe the model and its intended uses.
- Describe the relationship of the model to data, including the data for both inputs and corroboration.

- Describe how such data and other sources of information will be used to assess the ability of the model to meet its intended task.
- Describe all the elements of the evaluation plan by using an outline or diagram that shows how the elements relate to the model's life cycle.
- Describe the factors or events that might trigger the need for major model revisions or the circumstances that might prompt users to seek an alternative model. These can be fairly broad and qualitative.
- Identify the responsibilities, accountabilities, and resources needed to ensure implementation of the evaluation plan.

As stated above, the goal of model evaluation is to ensure model quality. At EPA, quality is defined by the Information Quality Guidelines (IQGs) (EPA 2002a). The IQGs apply to all information that EPA disseminates, including models, information from models, and input data (see Appendix C, Box C4: Definition of Quality). According to the IQGs, quality has three major components: integrity, utility, and objectivity. This chapter focuses on addressing the four questions listed above by evaluating the third component, objectivity — specifically, how to ensure the objectivity of information from models by considering their accuracy, bias, and reliability.

- Accuracy, as described in Section 2.4, is the closeness of a measured or computed value to its "true" value, where the "true" value is obtained with perfect information.
- Bias describes any systematic deviation between a measured (i.e., observed) or computed value and its "true" value. Bias is affected by faulty instrument calibration and other measurement errors, systematic errors during data collection, and sampling errors such as incomplete spatial randomization during the design of sampling programs.
- Reliability is the confidence that (potential) users have in a model and its outputs such that they are willing to use the model and accept its results (Sargent 2000). Specifically, reliability is a function of the model's performance record and its conformance to best available, practicable science.

This chapter describes principles, tools, and considerations for model evaluation throughout all stages of development and application. Section 4.2 presents a variety of qualitative and quantitative best practices for evaluating models. Section 4.3 discusses special considerations for evaluating proprietary models. Section 4.4 explains why retrospective analysis of models, conducted after a model has been applied, can be important to improve individual models and regulatory policies and to systematically enhance the overall modeling field. Finally, Section 4.5 describes how the evaluation process culminates in a decision whether to apply the model to decision making. Section 4.6 reviews the key recommendations from this chapter.

4.2 Best Practices for Model Evaluation

The four questions listed above address the soundness of the science underlying a model, the quality and quantity of available data, the degree of correspondence with observed conditions, and the appropriateness of a model for a given application. This guidance describes several "tools" or best practices to address these questions: peer review of models; QA project planning, including data quality assessment; model corroboration (qualitative and/or quantitative evaluation of a model's accuracy and predictive capabilities); and sensitivity and uncertainty analysis. These tools and practices include both qualitative and quantitative techniques:

- Qualitative assessments: Some of the uncertainty in model predictions may arise from sources whose uncertainty cannot be quantified. Examples are uncertainties about the theory underlying the model, the manner in which that theory is mathematically expressed to represent the environmental components, and the theory being modeled. Subjective evaluation of experts may be needed to determine appropriate values for model parameters and inputs that cannot be directly observed or measured (e.g., air emissions estimates). Qualitative assessments are needed for these sources of uncertainty. These assessments may involve expert elicitation regarding the system's behavior and comparison with model forecasts.
- Quantitative assessments: The uncertainty in some sources — such as some model parameters and some input data — can be estimated through quantitative assessments involving statistical uncertainty and sensitivity analyses. These types of analyses can also be used to quantitatively describe how model estimates of current conditions may be expected to differ from comparable field observations. However, since model predictions are not directly observed, special care is needed when quantitatively comparing model predictions with field data.

As discussed previously, model evaluation is an iterative process. Hence, these tools and techniques may be effectively applied throughout model development, testing, and application and should not be interpreted as sequential steps for model evaluation.

Model evaluation should always be conducted using a graded approach that is adequate and appropriate to the decision at hand (EPA 2001, 2002b). This approach recognizes that model evaluation can be modified to the circumstances of the problem at hand and that programmatic requirements are varied. For example, a screening model (a type of model designed to provide a "conservative" or risk-averse answer) that is used for risk management should undergo rigorous evaluation to avoid false negatives, while still not imposing unreasonable data-generation burdens (false positives) on the regulated community. Ideally, decision makers and modeling staff work together at the onset of new projects to identify the appropriate degree of model evaluation (see Section 3.1).

External circumstances can affect the rigor required in model evaluation. For example, when the likely result of modeling will be costly control strategies and associated controversy, more detailed model evaluation may be necessary. In these cases, many aspects of the modeling may come under close scrutiny, and the modeler must document the findings of the model evaluation process and be prepared to answer questions that will arise about the model. A deeper level of model evaluation may also be appropriate when modeling unique or extreme situations that have not been previously encountered.

Finally, as noted earlier, some assessments require the use of multiple, linked models. This linkage has implications for assessing uncertainty and applying the system of models. Each component model as well as the full system of integrated models must be evaluated.

Sections 4.2.1 and 4.2.2, on peer review of models and quality assurance protocols for input data, respectively, are drawn from existing guidance. Section 4.2.3, on model corroboration activities and the use of sensitivity and uncertainty analysis, provides new guidance for model evaluation (along with Appendix D).

Box 6: Examples of Life Cycle Model Evaluation

The value in evaluating a model from the conceptual stage through the use stage is illustrated in a multi-year project conducted by the Organization for Economic Cooperation and Development (OECD). The project sought to develop a screening model that could be used to assess the persistence and long-range transport potential of chemicals. To ensure its effectiveness, the screening model needed to be a consensus model that had been evaluated against a broad set of available models and data.

This project began at a 2001 workshop to set model performance and evaluation goals that would provide the foundation for subsequent model selection and development (OECD 2002). OECD then established an expert group in 2002. This group began its work by developing and publishing a guidance document on using multimedia models to estimate environmental persistence and long-range transport. From 2003 to 2004, the group compared and assessed the performance of nine available multimedia fate and transport models (Fenner et al. 2005; Klasmeier et al. 2006). The group then developed a parsimonious consensus model representing the minimum set of key components identified in the model comparison. They convened three international workshops to disseminate this consensus model and provide an ongoing model evaluation forum (Scheringer et al. 2006).

In this example, more than half the total effort was invested in the conceptual and model formulation stages, and much of the effort focused on performance evaluation. The group recognized that each model's life cycle is different, but noted that attention should be given to developing consensus-based approaches in the model concept and formulation stages. Conducting concurrent evaluations at these stages in this setting resulted in a high degree of buy-in from the various modeling groups.

4.2.1 Scientific Peer Review

Peer review provides the main mechanism for independent evaluation and review of environmental models used by the Agency. Peer review provides an independent, expert review of the evaluation in Section 4.1; therefore, its purpose is two-fold:

- To evaluate whether the assumptions, methods, and conclusions derived from environmental models are based on sound scientific principles.
- To check the scientific appropriateness of a model for informing a specific regulatory decision. (The latter objective is particularly important for secondary applications of existing models.)

Information from peer reviews is also helpful for choosing among multiple competing models for a specific regulatory application. Finally, peer review is useful to identify the limitations of existing models. Peer review is *not* a mechanism to comment on the *regulatory decisions* or policies that are informed by models (EPA 2000c).

Peer review charge questions and corresponding records for peer reviewers to answer those questions should be incorporated into the quality assurance project plan, developed during assessment planning (see Section 4.2.2, below). For example, peer reviews may focus on whether a model meets the objectives or specifications that were set as part of the quality assurance plan (see EPA 2002b) (see Section 3.1).

All models that inform *significant*[2] regulatory decisions are candidates for peer review (EPA 2000c, 1993) for several reasons:

- Model results will be used as a basis for major regulatory or policy/guidance decision making.
- These decisions likely involve significant investment of Agency resources.
- These decisions may have inter-Agency or cross-agency implications/applicability.

Existing guidance recommends that a new model should be scientifically peer-reviewed prior to its first application; for subsequent applications, the program manager should consider the scientific/technical complexity and/or the novelty of the particular circumstances to determine whether additional peer review is needed (EPA 1993). To conserve resources, peer review of "similar" applications should be avoided.

Models used for secondary applications (existing EPA models or proprietary models) will generally undergo a different type of evaluation than those developed with a specific regulatory information need in mind. Specifically, these reviews may deal more with uncertainty about the appropriate application of a model to a specific set of conditions than with the science underlying the model framework. For example, a project team decides to assess a water quality problem using WASP, a well-established water quality model framework. The project team determines that peer review of the model framework itself is not necessary, and the team instead conducts a peer review on their specific application of the WASP framework.

The following aspects of a model should be peer-reviewed to establish scientific credibility (SAB 1993a, EPA 1993):

- Appropriateness of input data.
- Appropriateness of boundary condition specifications.
- Documentation of inputs and assumptions.
- Applicability and appropriateness of selected parameter values.
- Documentation and justification for adjusting model inputs to improve model performance (calibration).
- Model application with respect to the range of its validity.
- Supporting empirical data that strengthen or contradict the conclusions that are based on model results.

To be most effective and maximize its value, external peer review should begin as early in the model *development* phase as possible (EPA 2000b). Because peer review involves significant time and resources, these allocations must be incorporated into components of the project planning and any

[2] Executive Order 12866 (58 FR 51735) requires federal agencies to determine whether a regulatory action is "significant" and therefore, subject to the requirements of the Executive Order, including review by the Office of Management and Budget. The Order defines "significant regulatory action" as one "that is likely to result in a rule that may: (1) Have an annual effect on the economy of $100 million or more or adversely affect in a material way the economy, a sector of the economy, productivity, competition, jobs, the environment, public health or safety, or State, local, or tribal governments or communities; (2) Create a serious inconsistency or otherwise interfere with an action taken or planned by another agency; (3) Materially alter the budgetary impacts of entitlements, grants, user fees, or loan programs or the rights and obligations of recipients thereof; or (4) Raise novel legal or policy issues arising out of legal mandates, the President's priorities, or the principles set forth in [the] Order." Section 2(f).

related contracts. Peer review in the early stages of model development can help evaluate the conceptual basis of models and potentially save time by redirecting misguided initiatives, identifying alternative approaches, or providing strong technical support for a potentially controversial position (SAB 1993a, EPA 1993). Peer review in the later stages of model development is useful as an independent external review of model code (i.e., model verification). External peer review of the *applicability* of a model to a particular set of conditions should be considered well in advance of any decision making, as it helps avoid inappropriate applications of a model for specific regulatory purposes (EPA 1993).

The peer review logistics are left to the discretion of the managers responsible for applying the model results to decision making. Mechanisms for accomplishing external peer review include (but are not limited to):

- Using an ad hoc panel of scientists.[3]
- Using an established external peer review mechanism such as the SAB
- Holding a technical workshop.[4]

Several sources provide guidance for determining the qualifications and number of reviewers needed for a given modeling project (SAB 1993a; EPA 2000c, 1993, 1994a). Key aspects are summarized in Appendix D of this guidance.

4.2.2 Quality Assurance Project Planning and Data Quality Assessment

Like peer review, data quality assessment addresses whether a model has been developed according to the principles of sound science. While some variability in data is unavoidable (see Section 4.2.3.1), adhering to the tenets of data quality assessment described in other Agency guidance[5] (Appendix D, Box D2: Quality Assurance Planning and Data Acceptance Criteria) helps minimize data uncertainty.

Well-executed QA project planning also helps ensure that a model performs the specified task, which addresses the fourth model evaluation question posed in Section 4.1. As discussed above, evaluating the degree to which a modeling project has met QA objectives is often a function of the external peer review process. The *Guidance for Quality Assurance Project Plans for Modeling* (EPA 2002b) provides general information about how to document quality assurance planning for modeling (e.g., specifications

[3] The formation and use of an ad hoc panel of peer reviewers may be subject to the Federal Advisory Committee Act (FACA). Compliance with FACA's requirements is summarized in Chapter Two of the *Peer Review Handbook,* "Planning a Peer Review" (EPA 2000c). Guidance on compliance with FACA may be sought from the Office of Cooperative Environmental Management. Legal questions regarding FACA may be addressed to the Cross-Cutting Issues Law Office in the Office of General Counsel.

[4] Note that a technical workshop held for peer review purposes is not subject to FACA *if the reviewers provide individual opinions.* [Note that there is no "one time meeting" exemption from FACA. The courts have held that even a single meeting can be subject to FACA.] An attempt to obtain group advice, whether it be consensus or majority-minority views, likely would trigger FACA requirements.

[5] Other guidance that can help ensure the quality of data used in modeling projects includes:

- *Guidance for the Data Quality Objectives Process,* a systematic planning process for environmental data collection (EPA 2000a).
- *Guidance on Choosing a Sampling Design for Environmental Data Collection,* on applying statistical sampling designs to environmental applications (EPA 2002c).
- *Guidance for Data Quality Assessment: Practical Methods for Data Analysis*, to evaluate the extent to which data can be used for a specific purpose (EPA 2000b).

or assessment criteria development, assessments of various stages of the modeling process; reports to management as feedback for corrective action; and finally the process for acceptance, rejection, or qualification of the output for use) to conform with EPA policy and acquisition regulations. Data quality assessments are a key component of the QA plan for models.

Both the quality and quantity (representativeness) of supporting data used to parameterize and (when available) corroborate models should be assessed during all relevant stages of a modeling project. Such assessments are needed to evaluate whether the available data are sufficient to support the choice of the model to be applied (question 2, Section 4.1), and to ensure that the data are sufficiently representative of the true system being modeled to provide meaningful comparison to observational data (question 3, Section 4.1).

4.2.3 Corroboration, Sensitivity Analysis, and Uncertainty Analysis

The question "How closely does the model approximate the real system of interest?" is unlikely to have a simple answer. In general, answering this question is not simply a matter of comparing model results and empirical data. As noted in Section 3.1, when developing and using an environmental model, modelers and decision makers should consider what degree of uncertainty is acceptable within the context of a specific model application. To do this, they will need to understand the uncertainties underlying the model. This section discusses three approaches to gaining this understanding:

- Model corroboration (Section 4.2.3.2), which includes all quantitative and qualitative methods for evaluating the degree to which a model corresponds to reality.
- Sensitivity analysis (Section 4.2.3.3), which involves studying how changes in a model's input values or assumptions affect its output or response.
- Uncertainty analysis (Section 4.2.3.3), which investigates how a model might be affected by the lack of knowledge about a certain population or the real value of model parameters.

Where practical, the recommended analyses should be conducted and their results reported in the documentation supporting the model. Section 4.2.3.1 describes and defines the various types of uncertainty, and associated concepts, inherent in the modeling process that model corroboration and sensitivity and uncertainty analysis can help assess.

4.2.3.1 Types of Uncertainty

Uncertainties are inherent in all aspects of the modeling process. Identifying those uncertainties that *significantly* influence model outcomes (either qualitatively or quantitatively) and communicating their importance is key to successfully integrating information from models into the decision making process. As defined in Chapter 3, uncertainty is the term used in this guidance to describe incomplete knowledge about specific factors, parameters (inputs), or models. For organizational simplicity, uncertainties that affect model quality are categorized in this guidance as:

- **Model framework uncertainty,** resulting from incomplete knowledge about factors that control the behavior of the system being modeled; limitations in spatial or temporal resolution; and simplifications of the system.

- **Model input uncertainty,** resulting from data measurement errors, inconsistencies between measured values and those used by the model (e.g., in their level of aggregation/averaging), and parameter value uncertainty.
- **Model niche uncertainty,** resulting from the use of a model outside the system for which it was originally developed and/or developing a larger model from several existing models with different spatial or temporal scales.

Box 7: Example of Model Input Uncertainty

The NRC's *Models in Environmental Regulatory Decision Making* provides a detailed example, summarized below, of the effect of model input uncertainty on policy decisions.

The formation of ozone in the lower atmosphere (troposphere) is an exceedingly complex chemical process that involves the interaction of oxides of nitrogen (NO_x), volatile organic compounds (VOCs), sunlight, and dynamic atmospheric processes. The basic chemistry of ozone formation was known in the early 1960s (Leighton 1961). Reduction of ozone concentrations generally requires controlling either or both NO_x and VOC emissions. Due to the nonlinearity of atmospheric chemistry, selection of the emission-control strategy traditionally relied on air quality models.

One of the first attempts to include the complexity of atmospheric ozone chemistry in the decision making process was a simple observation-based model, the so-called Appendix J curve (36 Fed. Reg. 8166 [1971]). The curve was used to indicate the percentage VOC emission reduction required to attain the ozone standard in an urban area based on peak concentration of photochemical oxidants observed in that area. Reliable NO_x data were virtually nonexistent at the time; Appendix J was based on data from measurements of ozone and VOC concentrations from six U.S. cities. The Appendix J curve was based on the hypothesis that reducing VOC emissions was the most effective emission-control path, and this conceptual model helped define legislative mandates enacted by Congress that emphasized controlling these emissions.

The choice in the 1970s to concentrate on VOC controls was supported by early results from models. Though new results in the 1980s showed higher-than-expected biogenic VOC emissions, EPA continued to emphasize VOC controls, in part because the schedule that Congress and EPA set for attaining the ozone ambient air quality standards was not conducive to reflecting on the basic elements of the science (Dennis 2002).

VOC reductions from the early 1970s to the early 1990s had little effect on ozone concentrations. Regional ozone models developed in the 1980s and 1990s suggested that controlling NO_x emissions was necessary in addition to, or instead of, controlling VOCs to reduce ozone concentrations (NRC 1991). The shift in the 1990s toward regulatory activities focusing on NO_x controls was partly due to the realization that historical estimates of emissions and the effectiveness of various control strategies in reducing emissions were not accurate. In other words, ozone concentrations had not been reduced as much as hoped over the past three decades, in part because emissions of some pollutants were much higher than originally estimated.

Regulations may go forward before science and models are perfected because of the desire to mitigate the potential harm from environmental hazards. In the case of ozone modeling, the model inputs (emissions inventories in this case) are often more important than the model science (description of atmospheric transport and chemistry in this case) and require as careful an evaluation as the evaluation of the model. These factors point to the potential synergistic role that measurements play in model development and application.

In reality, all three categories are interrelated. Uncertainty in the underlying model structure or model framework uncertainty is the result of incomplete scientific data or lack of knowledge about the factors

that control the behavior of the system being modeled. Model framework uncertainty can also be the result of simplifications needed to translate the conceptual model into mathematical terms as described in Section 3.3. In the scientific literature, this type of uncertainty is also referred to as structural error (Beck 1987), conceptual errors (Konikow and Bredehoeft 1992), uncertainties in the conceptual model (Usunoff et al. 1992), or model error/uncertainty (EPA 1997; Luis and McLaughlin 1992). Structural error relates to the mathematical construction of the algorithms that make up a model, while the conceptual model refers to the science underlying a model's governing equations. The terms "model error" and "model uncertainty" are both generally synonymous with model framework uncertainty.

Many models are developed iteratively to update their underlying science and resolve existing model framework uncertainty as new information becomes available. Models with long lives may undergo important changes from version to version. The MOBILE model for estimating atmospheric vehicle emissions, the CMAQ (Community Multi-scale Air Quality) model, and the QUAL2 water quality models are examples of models that have had multiple versions and major scientific modifications and extensions in over two decades of their existence (Scheffe and Morris 1993; Barnwell et al. 2004; EPA 1999c, as cited in NRC 2007).

When an appropriate model framework has been developed, the model itself may still be highly uncertain if the input data or database used to construct the application tool is not of sufficient quality. The quality of empirical data used for both model parameterization and corroboration tests is affected by both uncertainty and variability. This guidance uses the term "data uncertainty" to refer to the uncertainty caused by measurement errors, analytical imprecision, and limited sample sizes during data collection and treatment.

In contrast to data uncertainty, variability results from the inherent randomness of certain parameters, which in turn results from the heterogeneity and diversity in environmental processes. Examples of variability include fluctuations in ecological conditions, differences in habitat, and genetic variances among populations (EPA 1997). Variability in model parameters is largely dependent on the extent to which input data have been aggregated (both spatially and temporally). Data uncertainty is sometimes referred to as reducible uncertainty because it can be minimized with further study (EPA 1997). Accordingly, variability is referred to as irreducible because it can be better characterized and represented but not reduced with further study (EPA 1997).

A model's application niche is the set of conditions under which use of the model is scientifically defensible (EPA 1994b). Application niche uncertainty is therefore a function of the appropriateness of a model for use under a specific set of conditions. Application niche uncertainty is particularly important when (a) choosing among existing models for an application that lies outside the system for which the models were originally developed and/or (b) developing a larger model from several existing models with different spatial or temporal scales (Levins 1992).

The SAB's review of MMSOILS (Multimedia Contaminant Fate, Transport and Exposure Model) provides a good example of application niche uncertainty. The SAB questioned the adequacy of using a screening-level model to characterize situations where there is substantial subsurface heterogeneity or where non-aqueous phase contaminants are present (conditions differ from default values) (SAB 1993b). The SAB considered the MMSOILS model acceptable within its original application niche, but unsuitable for more heterogeneous conditions.

4.2.3.2 Model Corroboration

> *The interdependence of models and measurements is complex and iterative for several reasons. Measurements help to provide the conceptual basis of a model and inform model development, including parameter estimation. Measurements are also a critical tool for corroborating model results. Once developed, models can derive priorities for measurements that ultimately get used in modifying existing models or in developing new ones. Measurement and model activities are often conducted in isolation...Although environmental data systems serve a range of purposes, including compliance assessment, monitoring of trends in indicators, and basic research performance, the importance of models in the regulatory process requires measurements and models to be better integrated. Adaptive strategies that rely on iterations of measurements and modeling, such as those discussed in the 2003 NRC report titled* Adaptive Monitoring and Assessment for the Comprehensive Everglades Restoration Plan, *provide examples of how improved coordination might be achieved.*
>
> *— NRC Committee on Models in the Regulatory Decision Process (NRC 2007)*

Model corroboration includes all quantitative and qualitative methods for evaluating the degree to which a model corresponds to reality. The rigor of these methods varies depending on the type and purpose of the model application. Quantitative model corroboration uses statistics to estimate how closely the model results match measurements made in the real system. Qualitative corroboration activities may include expert elicitation to obtain beliefs about a system's behavior in a data-poor situation. These corroboration activities may move model forecasts toward consensus.

For newly developed model frameworks or untested mathematical processes, formal corroboration procedures may be appropriate. Formal corroboration may involve formulation of hypothesis tests for model acceptance, tests on datasets independent of the calibration dataset, and quantitative testing criteria. In many cases, collecting independent datasets for formal model corroboration is extremely costly or otherwise unfeasible. In such circumstances, model evaluation may be appropriately conducted using a combination of other evaluation tools discussed in this section.

Robustness is the capacity of a model to perform equally well across the full range of environmental conditions for which it was designed (Reckhow 1994; Borsuk et al. 2002). The degree of similarity among datasets available for calibration and corroboration provides insight into a model's robustness. For example, if the dataset used to corroborate a model is identical or statistically similar to the dataset used to calibrate the model, then the corroboration exercise has provided neither an independent measure of the model's performance nor insight into the model's robustness. Conversely, when corroboration data are significantly different from calibration data, the corroboration exercise provides a measure of both model performance and robustness.

Quantitative model corroboration methods are recommended for choosing among multiple models that are available for the same application. In such cases, models may be ranked on the basis of their statistical performance in comparison to the observational data (e.g., EPA 1992). EPA's Office of Air and Radiation evaluates models in this manner. When a single model is found to perform better than others in a given category, OAR recommends it in the *Guidelines on Air Quality Models* as a preferred model for

application in that category (EPA 2003a). If models perform similarly, then the preferred model is selected based on other factors, such as past use, public familiarity, cost or resource requirements, and availability.

Box 8: Example: Comparing Results from Models of Varying Complexity
(From Box 5-4 in NRC's *Models in Environmental Regulatory Decision Making*)

The Clean Air Mercury Rule[6] requires industry to reduce mercury emissions from coal-fired power plants. A potential benefit is the reduced human exposure and related health impacts from methylmercury that may result from reduced concentrations of this toxin in fish. Many challenges and uncertainties affect assessment of this benefit. In its assessment of the benefits and costs of this rule, EPA used multiple models to examine how changes in atmospheric deposition would affect mercury concentrations in fish, and applied the models to assess some of the uncertainties associated with the model results (EPA 2005).

EPA based its national-scale benefits assessment on results from the mercury maps (MMaps) model. This model assumes a linear, steady-state relationship between atmospheric deposition of mercury and mercury concentrations in fish, and thus assumes that a 50% reduction in mercury deposition rates results in a 50% decrease in fish mercury concentrations. In addition, MMaps assumes instantaneous adjustment of aquatic systems and their ecosystems to changes in deposition — that is, no time lag in the conversion of mercury to methylmercury and its bioaccumulation in fish. MMaps also does not deal with sources of mercury other than those from atmospheric deposition. Despite those limitations, the Agency concluded that no other available model was capable of performing a national-scale assessment.

To further investigate fish mercury concentrations and to assess the effects of MMaps' assumptions, EPA applied more detailed models, including the spreadsheet-based ecological risk assessment for the fate of mercury (SERAFM) model, to five well-characterized ecosystems. Unlike the steady-state MMaps model, SERAFM is a dynamic model which calculates the temporal response of mercury concentrations in fish tissues to changes in mercury loading. It includes multiple land-use types for representing watershed loadings of mercury through soil erosion and runoff. SERAFM partitions mercury among multiple compartments and phases, including aqueous phase, abiotic participles (for example, silts), and biotic particles (for example, phytoplankton). Comparisons of SERAFM's predictions with observed fish mercury concentrations for a single fish species in four ecosystems showed that the model under-predicted mean concentrations for one water body, over-predicted mean concentrations for a second water body, and accurately predicted mean concentrations for the other two. The error bars for the observed fish mercury concentrations in these four ecosystems were large, making it difficult to assess the models' accuracy. Modeling the four ecosystems also showed how the assumed physical and chemical characteristics of the specific ecosystem affected absolute fish mercury concentrations and the length of time before fish mercury concentrations reached steady state.

Although EPA concluded that the best available science supports the assumption of a linear relationship between atmospheric deposition and fish mercury concentrations for broad-scale use, the more detailed ecosystem modeling demonstrated that individual ecosystems were highly sensitive to uncertainties in model parameters. The Agency also noted that many of the model uncertainties could not be quantified. Although the case studies covered the bulk of the key environmental characteristics, EPA found that extrapolating the individual ecosystem case studies to account for the variability in ecosystems across the country indicated that those case studies might not represent extreme conditions that could influence how atmospheric mercury deposition affected fish mercury concentrations in

[6] On February 8, 2008, the U.S. Court of Appeals for the District of Columbia Circuit vacated the Clean Air Mercury Rule. The DC Circuit's vacatur of this rule was unrelated to the modeling conducted in support of the rule.

a water body.

This example illustrates the usefulness of investigating a variety of models at varying levels of complexity. A hierarchical modeling approach, such as that used in the mercury analysis, can provide justification for simplified model assumptions or potentially provide evidence for a consistent bias that would negate the assumption that a simple model is appropriate for broad-scale application.

4.2.3.3 Sensitivity and Uncertainty Analysis

Sensitivity analysis is the study of how a model's response can be apportioned to changes in model inputs (Saltelli et al. 2000a). Sensitivity analysis is recommended as the principal evaluation tool for characterizing the most and least important sources of uncertainty in environmental models.

Uncertainty analysis investigates the lack of knowledge about a certain population or the real value of model parameters. Uncertainty can sometimes be reduced through further study and by collecting additional data. EPA guidance (e.g., EPA 1997) distinguishes uncertainty analysis from methods used to account for variability in input data and model parameters. As mentioned earlier, variability in model parameters and input data can be better characterized through further study but is usually not reducible (EPA 1997).

Although sensitivity and uncertainty analysis are closely related, sensitivity is algorithm-specific with respect to model "variables" and uncertainty is parameter-specific. Sensitivity analysis assesses the "sensitivity" of the model to specific parameters and uncertainty analysis assesses the "uncertainty" associated with parameter values. Both types of analyses are important to understand the degree of confidence a user can place in the model results. Recommended techniques for conducting uncertainty and sensitivity analysis are discussed in Appendix D.

The NRC committee pointed out that uncertainty analysis for regulatory environmental modeling involves not only analyzing uncertainty, but also communicating the uncertainties to policy makers. To facilitate communication of model uncertainty, the committee recommends using hybrid approaches in which unknown quantities are treated probabilistically *and* explored in scenario-assessment mode by decision makers through a range of plausible values. The committee further acknowledges (NRC 2007) that:

> *Effective uncertainty communication requires a high level of interaction with the relevant decision makers to ensure that they have the necessary information about the nature and sources of uncertainty and their consequences. Thus, performing uncertainty analysis for environmental regulatory activities requires extensive discussion between analysts and decision makers.*

4.3 Evaluating Proprietary Models

This guidance defines proprietary models as those computer models for which the source code is not universally shared. To promote the transparency with which decisions are made, EPA prefers using non-proprietary models when available. However, the Agency acknowledges there will be times when the use of proprietary models provides the most reliable and best-accepted characterization of a system.

D.4.4.4 Summary of Performance

As an example of overall summary of performance, we will discuss a procedure constructed using the scheme introduced by Cox and Tikvart (1990) as a template. The design for statistically summarizing model performance over several regimes is envisioned as a five-step procedure.

1. Form a replicate sample using concurrent sampling of the observed and modeled values for each regime. Concurrent sampling associates results from all models with each observed value, so that selection of an observed value automatically selects the corresponding estimates by all models.
2. Compute the average of observed and modeled values for each regime.
3. Compute the normalized mean square error, NMSE, using the computed regime averages, and store the value of the NMSE computed for this pass of the bootstrap sampling.
4. Repeat steps 1 through 3 for all bootstrap sampling passes (typically of order 500).
5. Implement the procedure described in ASTM D 6589 (ASTM 2000) to detect which model has the lowest computed NMSE value (call this the "base" model) and which models have NMSE values that are significantly different from the "base" model.

In the Cox and Tikvart (1990) analysis, the data were sorted into regimes (defined in terms of Pasquill stability category and low/high wind speed classes), and bootstrap sampling was used to develop standard error estimates on the comparisons. The performance measure was the robust highest concentration (computed from the raw observed cumulative frequency distribution), which is a comparison of the highest concentration values (maxima), which most models do not contain the physics to simulate. This procedure can be improved if intensive field data are used and the performance measure is the NMSE computed from the modeled and observed regime averages of centerline concentration values as a function of stability along each downwind arc, where each regime is a particular distance downwind for a defined stability range.

The data demands are much greater for using regime averages than for using individual concentrations. Procedures that analyze groups (regimes) of data include intensive tracer field studies, with a dense receptor network, and many experiments. Whereas, Cox and Tikvart (1990) devised their analysis to make use of very sparse receptor networks having one or more years of sampling results. With dense receptor networks, attempts can be made to compare average modeled and "observed" centerline concentration values, but only a few of these experiments have sufficient data to allow stratification of the data into regimes for analysis. With sparse receptor networks, there are more data for analysis, but there is insufficient information to define the observed maxima relative to the dispersing plume's center of mass. Thus, there is uncertainty as to whether or not the observed maxima are representative of centerline concentration values. It is not obvious that the average of the *n* (say 25) observed maximum hourly concentration values (for a particular distance downwind and narrowly defined stability range) is the ensemble average centerline concentration the model is predicting. In fact, one might anticipate that the average of the *n* maximum concentration values is likely to be higher than the ensemble average of the centerline concentration. Thus the testing procedure outlined by Cox and Tikvart (1990) may favor selection of poorly formed models that routinely underestimate the lateral diffusion (and thereby overestimate the plume centerline concentration). This in turn, may bias such models' ability to characterize concentration patterns for longer averaging times.

It is therefore concluded that once a set of "best-performing models" has been selected from an evaluation using intensive field data that tests a model's ability to predict the average characteristics to be seen in the observed concentration patterns, evaluations using sparse networks are seen as useful extensions to further explore the performance of well-formulated models for other environs and purposes.

D.5 Sensitivity Analysis

This section provides a broad overview of uncertainty and sensitivity analyses and introduces various methods used to conduct the latter. A table at the end of this section summarizes these methods' primary features and citations to additional resources for computational detail.

D.5.1 Introducing Sensitivity Analyses and Uncertainty Analysis

A model approximates reality in the face of scientific uncertainties. Section 4.1.3.1 identifies and defines various sources of model uncertainty. External peer reviewers of EPA models have consistently recommended that EPA communicate this uncertainty through uncertainty analysis and sensitivity analysis, two related disciplines. Uncertainty analysis investigates the effects of lack of knowledge or potential errors of model inputs (e.g., the "uncertainty" associated with parameter values); when combined with sensitivity analysis, it allows a model user to be more informed about the confidence that can be placed in model results. Sensitivity analysis measures the effect of changes in input values or assumptions (including boundaries and model functional form) on the outputs (Morgan and Henrion 1990); it is the study of how uncertainty in a model output can be systematically apportioned to different sources of uncertainty in the model input (Beck et al. 1994). By investigating the "relative sensitivity" of model parameters, a user can become knowledgeable of the relative importance of parameters in the model.

Consider a model represented as a function *f*, with inputs x_1 and x_2, and with output *y*, such that $y = f(x_1,x_2)$. Figure D.5.1 schematically depicts how uncertainty analysis and sensitivity analysis would be conducted for this model. Uncertainty analysis would be conducted by determining how *y* responds to variation in inputs x_1 and x_2, the graphic depiction of which is referred to as the model's response surface. Sensitivity analysis would be conducted by apportioning the respective contributions of x_1 and x_2 to changes in *y*. The schematic should *not* be construed to imply that uncertainty analysis and sensitivity analysis are sequential events. Rather, they are generally conducted by trial and error, with each type of analysis informing the other. Indeed, in practice, the distinction between these two related disciplines may be irrelevant. For purposes of clarity, the remainder of this appendix will refer exclusively to sensitivity analysis.

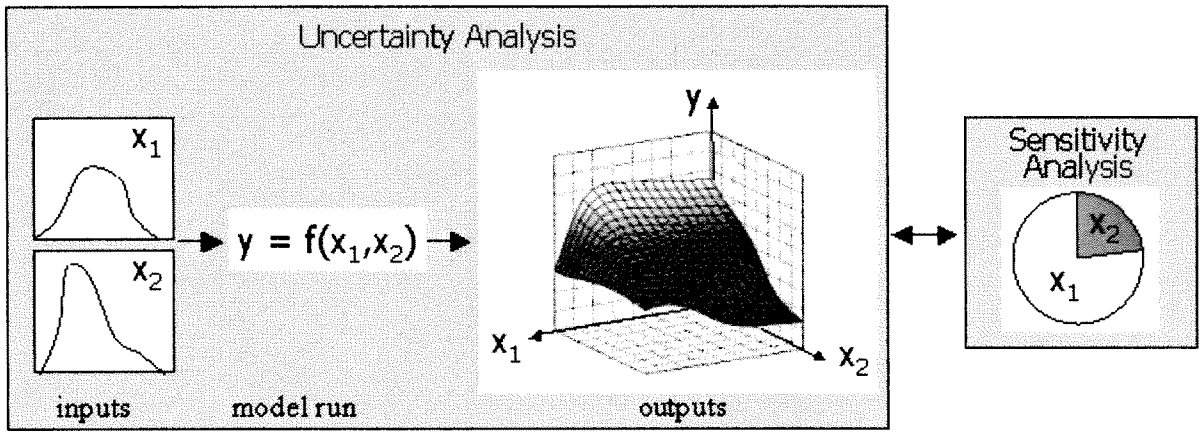


Figure D.5.1. Uncertainty and sensitivity analyses. Uncertainty analysis investigates the effects of lack of knowledge or potential errors of model inputs. Sensitivity analysis evaluates the respective contributions of inputs x_1 and x_2 to output *y*.

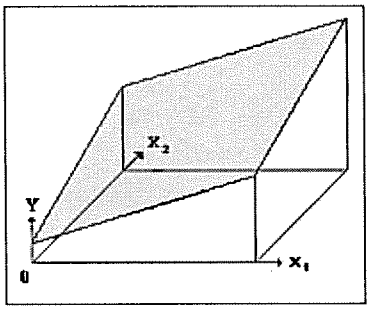
D.5.2 Sensitivity Analysis and Computational Complexity

Choosing the appropriate uncertainty analysis/sensitivity analysis method is often a matter of trading off between the amount of information one wants from the analyses and the computational difficulties of the analyses. These computational difficulties are often inversely related to the number of assumptions one is willing or able to make about the shape of a model's response surface.

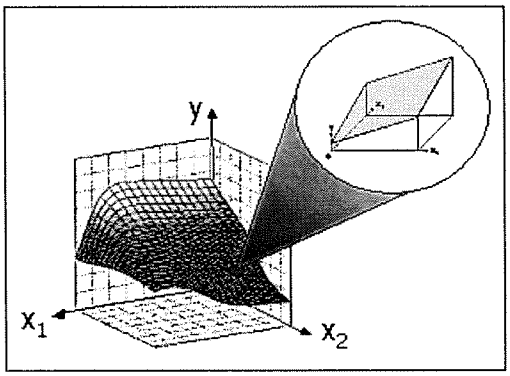
Consider once again a model represented as a function *f*, with inputs x_1 and x_2 and with output *y*, such that $y = f(x_1,x_2)$. *Sensitivity* measures how output changes with respect to an input. This is a straightforward enough procedure with differential analysis if the analyst:

- Can assume that the model's response surface is a hyperplane, as in Figure D.5.2(1);
- Accepts that the results apply only to specific points on the response surface and that these points are monotonic first order, as in Figure D.5.2 (2);[10] or
- Is unconcerned about interactions among the input variables.

Otherwise, sensitivity analysis may be more appropriately conducted using more intensive computational methods.



(1)



(2)

Figure D.5.2. It's hyperplane and simple. (1) A model response surface that is a hyperplane can simplify sensitivity analysis computations. (2) The same computations can also be used for other response surfaces, but only as approximations around a single locus.

This guidance suggests that, depending on assumptions underlying the model, the analyst should use non-intensive sensitivity analysis techniques to initially identify those inputs that generate the most sensitivity, then apply more intensive methods to this smaller subset of inputs. It may therefore be useful to categorize the various sensitivity analysis techniques into methods that (a) can be quickly used to screen for the more important input factors; (b) are based on differential analyses; (c) are based on sampling; and (d) are based on variance methods.

D.5.3 Screening Tools

D.5.3.1 Tools That Require No Model Runs

Cullen and Frey (1999) suggest that summary statistics measuring input uncertainty can serve as preliminary screening tools without additional model runs (and if the models are simple and linear), indicating proportionate contributions to output uncertainty:

- *Coefficient of variation.* The coefficient of variation is the standard deviation normalized to the mean (σ/μ) in order to reduce the possibility that inputs that take on large values are given undue importance.
- *Gaussian approximation.* Another approach to apportioning input variance is Gaussian approximation. Using this method, the variance of a model's output is estimated as the sum of the variances of the inputs (for additive models) or the sum of the variances of the log-transformed inputs (for multiplicative models), weighted by the squares on any constants which may be multiplied by the inputs as they occur in the model.

D.5.3.2 Scatterplots

Cullen and Frey (1999) suggest that a high correlation between an input and an output variable may indicate substantial dependence of the variation in output and the variation of the input. A simple, visual

[10] Related to this issue are the terms "local sensitivity analysis" and "global sensitivity analysis." The former refers to sensitivity analysis conducted around a nominal point of the response surface, while the latter refers to sensitivity analysis across the entire surface.

assessment of the influence of an input on the output is therefore possible using scatterplots, with each plot posing a selected input against the output, as in Figure D.5.3.

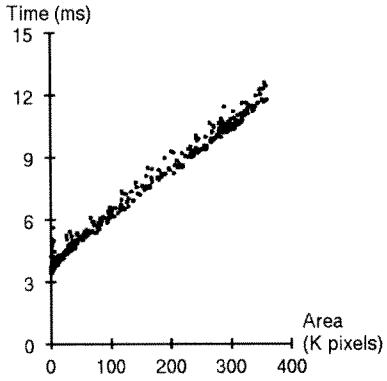


Figure D.5.3. Correlation as indication of input effect. The high correlation between the input variable area and the output variable time (holding all other variables fixed) is an indication of the possible effect of area's variation on the output.

D.5.3.3 Morris's OAT

The key concept underlying one-at-a-time (OAT) sensitivity analyses is to choose a base case of input values and to perturb each input variable by a given percentage away from the base value while holding all other input variables constant. Most OAT sensitivity analysis methods yield *local* measures of sensitivity (see footnote 9) that depend on the choice of base case values. To avoid this bias, Saltelli et al. (2000b) recommend using Morris's OAT for screening purposes because it is a *global* sensitivity analysis method — it entails computing a number of local measures (randomly extracted across the input space) and then taking their average.

Morris's OAT provides a measure of the importance of an input factor in generating output variation, and while it does not quantify interaction effects, it does provide an indication of the presence of interaction. Figure D.5.4 presents the results that one would expect to obtain from applying Morris's OAT (Cossarini et al. 2002). Computational methods for this technique are described in Saltelli et al. 2000b.

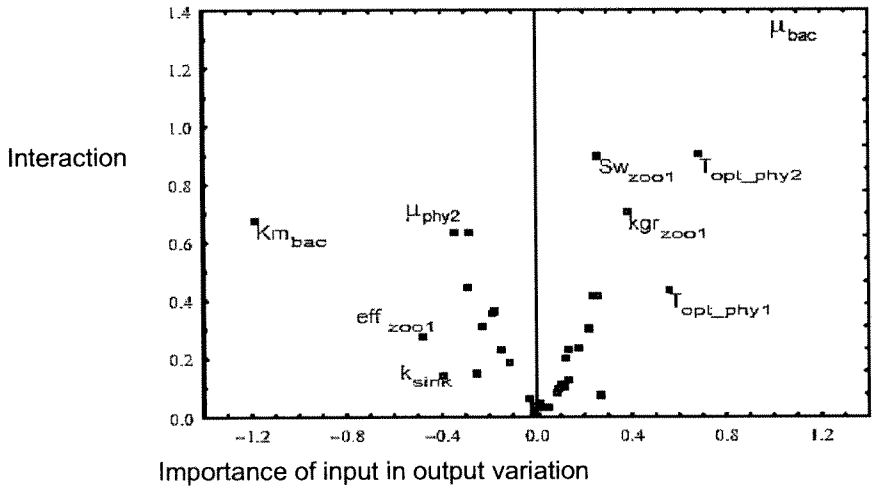


Figure D.5.4. An application of Morris's OAT. Cossarini et al. (2002) investigated the influence of various ecological factors on energy flow through a food web. Their sensitivity analysis indicated that maximum bacteria growth and bacteria mortality (μ_{bac} and Km_{bac}, respectively) have the largest (and opposite) effects on energy flow, as indicated by their values on the horizontal axis. These effects, as indicated by their values on the vertical axis, resulted from interactions with other factors.

D.5.4 Methods Based on Differential Analysis

As noted previously, differential analyses may be used to analyze sensitivity if the analyst is willing either to assume that the model response surface is hyperplanar or to accept that the sensitivity analysis results are local and that they are based on hyperplanar approximations tangent to the response surface at the nominal scenario (Morgan and Henrion 1990; Saltelli et al. 2000b).

Differential analyses entail four steps. First, select base values and ranges for input factors. Second, using these input base values, develop a Taylor series approximation to the output. Third, estimate uncertainty in output in terms of its expected value and variance using variance propagation techniques. Finally, use the Taylor series approximations to estimate the importance of individual input factors (Saltelli et al. 2000b). Computational methods for this technique are described in Morgan and Henrion 1990.

D.5.5 Methods Based on Sampling

One approach to estimating the impact of input uncertainties is to repeatedly run a model using randomly sampled values from the input space. The most well-known method using this approach is Monte Carlo analysis. In a Monte Carlo simulation, a model is run repeatedly. With each run, different input values are drawn randomly from the probability distribution functions of each input, thereby generating multiple output values (Morgan and Henrion 1990; Cullen and Frey 1999). One can view a Monte Carlo simulation as a process through which multiple scenarios generate multiple output values; although each execution of the model run is deterministic, the set of output values may be represented as a cumulative distribution function and summarized using statistical measures (Cullen and Frey 1999).

EPA proposes several best principles of good practice for the conduct of Monte Carlo simulations (EPA 1997). They include the following:

- Conduct preliminary sensitivity analyses to identify significant model components and input variables that make important contributions to model uncertainty.
- When deciding upon a probability distribution function (PDF) for input variables, consider the following questions: Is there any mechanistic basis for choosing a distributional family? Is the PDF likely to be dictated by physical, biological, or other properties and mechanisms? Is the variable

discrete or continuous? What are the bounds of the variable? Is the PDF symmetric or skewed, and if skewed, in which direction?
- Base the PDF on empirical, representative data.
- If expert judgment is used as the basis for the PDF, document explicitly the reasoning underlying this opinion.
- Discuss the presence or absence of covariance among the input variables, which can significantly affect the output.

The preceding points merely summarize some of the main points raised in EPA's Guidance on Monte Carlo Analysis. That document should be consulted for more detailed guidance. Conducting Monte Carlo analysis may be problematic for models containing a large number of input variables. Fortunately, there are several approaches to dealing with this problem:

- *Brute force approach.* One approach is to increase sheer computing power. For example, EPA's ORD is developing a Java-based tool that facilitates Monte Carlo analyses across a cluster of PCs by harnessing the computing power of multiple workstations to conduct multiple runs for a complex model (Babendreier and Castleton 2002).
- *Smaller, structured trials.* The value of Monte Carlo lies not in the randomness of sampling, but in achieving representative properties of sets of points in the input space. Therefore, rather than sampling data from entire input space, computations may be through *stratified sampling* by dividing the input sample space into strata and sampling from within each stratum. A widely used method for stratified sampling is *Latin hypercube sampling,* comprehensively described in Cullen and Frey 1999.
- *Response surface model surrogate.* The analyst may also choose to conduct Monte Carlo not on the complex model directly, but rather on a response surface representation of it. The latter is a simplified representation of the relationship between a selected number of model outputs and a selected number of model inputs, with all other model inputs held at fixed values (Morgan and Henrion 1990; Saltelli et al. 2000b).

D.5.6 Methods Based on Variance

Consider once again a model represented as a function f, with inputs x_1 and x_2 and with output y, such that $y = f(x_1,x_2)$. The input variables are affected by uncertainties and may take on any number of possible values. Let X denote an input vector randomly chosen from among all possible values for x_1 and x_2. The output y for a given X can also be seen as a realization of a random variable Y. Let $E(Y|X)$ denote the expectation of Y conditional on a fixed value of X. If the total variation in y is matched by the variability in $E(Y|X)$ as x_1 is allowed to vary, this is an indication that variation in x_1 significantly affects y.

The variance-based approaches to sensitivity analysis are based on the estimation of what fraction of total variation of y is attributable to variability in $E(Y|X)$ as a subset of input factors are allowed to vary. Three methods for computing this estimation (correlation ratio, Sobol, and Fourier amplitude sensitivity test) are featured in Saltelli et al. 2000b.

D.5.7 Which Method to Use?

A panel of experts was recently assembled to review various sensitivity analysis methods. The panel refrained from explicitly recommending a "best" method and instead developed a list of attributes for preferred sensitivity analysis methods. The panel recommended that methods should preferably be able to deal with a model regardless of assumptions about a model's linearity and additivity, consider interaction effects among input uncertainties, cope with differences in the scale and shape of input PDFs, cope with differences in input spatial and temporal dimensions, and evaluate the effect of an input while all other inputs are allowed to vary as well (Frey 2002; Saltelli 2002). Of the various methods discussed above, only those based on variance (Section D.5.6) are characterized by these attributes. When one or more of the criteria are not important, the other tools discussed in this section will provide a reasonable sensitivity assessment.

As mentioned earlier, choosing the most appropriate sensitivity analysis method will often entail a trade-off between computational complexity, model assumptions, and the amount of information needed from

the sensitivity analysis. As an aid to sensitivity analysis method selection, the table below summarizes the features and caveats of the methods discussed above.

Method	Features	Caveats	Reference
Screening methods	May be conducted independent of model run	Potential for significant error if model is non-linear	Cullen and Frey 1999, pp. 247-8.
Morris's one-at-a-time	Global sensitivity analysis	Indicates, but does not quantify interactions	Saltelli et al. 2000b, p. 68.
Differential analyses	Global sensitivity analysis for linear model; local sensitivity analysis for nonlinear model	No treatment of interactions among inputs Assumes linearity, monotonicity, and continuity	Cullen and Frey 1999, pp. 186-94. Saltelli et al. 2000b, pp. 183-91
Monte Carlo analyses	Intuitive No assumptions regarding response surface	Depending on number of input variables, may be time-consuming to run, but methods to simplify are available May rely on assumptions regarding input PDFs	Cullen and Frey 1999, pp. 196-237 Morgan and Henrion 1990, pp. 198-216.
Variance-based	Robust and independent of model assumptions Addresses interactions	May be computationally difficult.	Saltelli et al. 2000b, pp. 167-97

D.6 Uncertainty Analysis

D.6.1 Model Suitability

An evaluation of model suitability to resolve application niche uncertainty (Section 4.1.3.1) should precede any evaluation of data uncertainty and model performance. The extent to which a model is suitable for a proposed application depends on:

- Mapping of model attributes to the problem statement
- The degree of certainty needed in model outputs
- The amount of reliable data available or resources available to collect additional data
- Quality of the state of knowledge on which the model is based
- Technical competence of those undertaking simulation modeling

Appropriate data should be available before any attempt is made to apply a model. A model that needs detailed, precise input data should not be used when such data are unavailable.

D.6.2 Data Uncertainty

There are two statistical paradigms that can be adopted to summarize data. The first employs classical statistics and is useful for capturing the most likely or "average" conditions observed in a given system. This is known as the "frequentist" approach to summarizing model input data. Frequentist statistics rely on measures of central tendency (median, mode, mean values) and represent uncertainty as the deviation from these metrics. A frequentist or "deterministic" model produces a single set of solutions for each model run. In contrast, the alternate statistical paradigm employs a probabilistic framework, which summarizes data according to their "likelihood" of occurrence. Input data are represented as distributions rather than a single numerical value and models outputs capture a range of possible values.

The classical view of probability defines the probability of an event occurring by the value to which the long run frequency of an event or quantity converges as the number of trials increases (Morgan and Henrion 1990). Classical statistics relies on measures of central tendency (mean, median, mode) to